

SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing

Junyi Ao^{1,2,*}, Rui Wang^{3,*}, Long Zhou^{4,*}, Chengyi Wang⁴, Shuo Ren⁴,
Yu Wu⁴, Shujie Liu⁴, Tom Ko¹, Qing Li², Yu Zhang^{1,5}, Zhihua Wei³,
Yao Qian⁴, Jinyu Li⁴, Furu Wei⁴

¹Department of Computer Science and Engineering,
Southern University of Science and Technology

²Department of Computing, The Hong Kong Polytechnic University

³Department of Computer Science and Technology, Tongji University

⁴Microsoft ⁵Peng Cheng Laboratory

Abstract

Motivated by the success of T5 (Text-To-Text Transfer Transformer) in pre-trained natural language processing models, we propose a unified-modal SpeechT5 framework that explores the **encoder-decoder pre-training for self-supervised speech/text representation learning**. The SpeechT5 framework consists of a shared encoder-decoder network and six modal-specific (speech/text) pre/post-nets. After preprocessing the input speech/text through the pre-nets, the shared encoder-decoder network models the sequence-to-sequence transformation, and then the post-nets generate the output in the speech/text modality based on the output of the decoder. Leveraging large-scale unlabeled speech and text data, we pre-train SpeechT5 to learn a unified-modal representation, hoping to improve the modeling capability for both speech and text. To align the textual and speech information into this unified semantic space, we propose a cross-modal vector quantization approach that randomly mixes up speech/text states with latent units as the interface between encoder and decoder. Extensive evaluations show the superiority of the proposed SpeechT5 framework on a wide variety of spoken language processing tasks, including automatic speech recognition, speech synthesis, speech translation, voice conversion, speech enhancement, and speaker identification. We release our code and model at <https://github.com/microsoft/SpeechT5>.

1 Introduction

Starting with ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), substantial work has shown that pre-trained models can significantly improve in various natural language processing (NLP) tasks

*Equal contribution. Work is done by the first two authors during internship at Microsoft Research Asia. Correspondence to: Long Zhou (lozhou@microsoft.com)

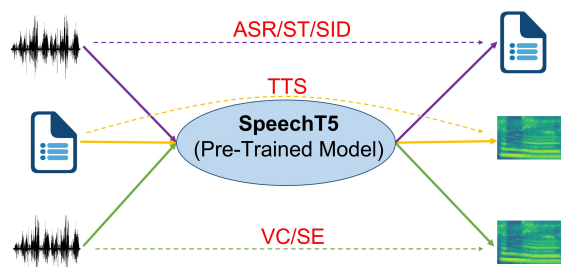


Figure 1: An illustration of the SpeechT5 framework, which treats spoken language processing tasks as a speech/text to speech/text format, including automatic speech recognition (ASR), speech translation (ST), speech identification (SID), text to speech (TTS), voice conversion (VC), and speech enhancement (SE).

(Radford et al., 2019; CONNEAU and Lample, 2019; Yang et al., 2019; Dong et al., 2019; Lewis et al., 2020). Following the pre-training techniques in NLP, self-supervised speech representation learning has also been investigated and shown promising results, benefiting from richly learned representations (Chung and Glass, 2018; Chuang et al., 2020; Song et al., 2019; Baevski et al., 2020; Wang et al., 2021; Hsu et al., 2021; Chung et al., 2021a), such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021).

However, previous speech pre-training work suffers from two problems: (1) most of them learn the speech representation with only unlabeled speech data but ignore the importance of textual data to spoken language tasks (e.g., automatic speech recognition) which require the modality transformation; (2) most of these models solely rely on a pre-trained speech encoder for various downstream tasks, leaving the decoder not pre-trained for the sequence-to-sequence generation tasks. How to design a unified encoder-decoder model that can take advantage of both unlabeled speech and text data to improve various spoken language processing tasks is not well explored.

Inspired by the T5 method (Raffel et al., 2020),

we attempt to formulate each spoken language processing task as a **speech/text to speech/text** problem via an encoder-decoder framework, which enables us to use the same pre-trained model with bimodal data across diverse tasks, as shown in Figure 1. To achieve this, we propose a unified-modal pre-training framework, SpeechT5, containing an encoder-decoder backbone network and **modal-specific pre/post-nets**. With the pre-nets, the **input speech/text is embedded in a shared space**, and the encoder-decoder backbone network models the sequence-to-sequence conversion, from which the **model-specific post-nets generate the speech/text output**. Particularly, SpeechT5 is mainly pre-trained with a denoising sequence-to-sequence method by leveraging large-scale unlabeled text and speech corpus. To align the textual and acoustic information into a unified semantic space, the proposed SpeechT5 model (1) maps text and speech representations into a shared vector quantization space, and (2) randomly mixes up the quantized latent representations and the contextual states, which can better guide the quantizer to learn the cross-modal features.

We fine-tune SpeechT5 on a wide variety of downstream spoken language processing tasks, including automatic speech recognition (ASR), text-to-speech (TTS), speech translation (ST), voice conversion (VC), speech enhancement (SE), and speaker identification (SID). Massive experiments show that the proposed SpeechT5 model achieves a significant improvement on these spoken language processing tasks compared with the state-of-the-art baselines. Specifically, the proposed SpeechT5 outperforms wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) with the BASE model on the ASR task and also performs better than the state-of-the-art voice Transformer network (Huang et al., 2021) on the VC task. Besides, SpeechT5 is significantly superior to SpeechNet (Chen et al., 2021b) and pre-trained models from SUPERB (Yang et al., 2021) and achieves the state-of-the-art performance (i.e., 96.49%) on the SID task. We further provide an empirical comparison of the pre-training tasks and modules, and the ablation study demonstrates the effectiveness of the proposed joint speech-text pre-training method.

The contributions of this paper are summarized as follows.

- To the best of our knowledge, this is the first work to investigate a unified encoder-decoder

framework for various spoken language processing tasks.

- We propose a cross-modal vector quantization approach, which learns the implicit alignment between acoustic and textual representation with large-scale unlabeled speech and text data.
- Extensive experiments on spoken language processing tasks demonstrate the effectiveness and superiority of the proposed SpeechT5 model.

2 SpeechT5

In this section, we propose SpeechT5, a unified-modal framework for learning joint contextual representations for speech and text data via a shared encoder-decoder structure.

2.1 Model Architecture

Figure 2(a) shows the model architecture of the proposed SpeechT5 model. It consists of an encoder-decoder module and six modal-specific pre/post-nets. The pre-nets convert the input speech $\mathbf{X}^s \in \mathcal{D}^s$ or text $\mathbf{X}^t \in \mathcal{D}^t$ to a unified space of hidden representations and then feed them into the shared encoder-decoder to perform the sequence-to-sequence conversion. Finally, the post-nets generate the output in the speech or text modality, based on the decoder output.

Input/Output Representations To train a single model for a diverse set of spoken language processing tasks, we formulate them as “speech/text to speech/text” tasks, where the model is fed with speech/text as the input and generates the corresponding output in the speech/text format. Specifically, a text is split into a sequence of characters $\mathbf{X}^t = (\mathbf{x}_1^t, \dots, \mathbf{x}_{N^t}^t)$ as the input and output. For speech modality, the raw waveform $\mathbf{X}^s = (\mathbf{x}_1^s, \dots, \mathbf{x}_{N^s}^s)$ is used as the input, and a sequence of the log Mel-filterbank features $\mathbf{X}^f = (\mathbf{x}_1^f, \dots, \mathbf{x}_{N^f}^f)$ extracted from raw audio using librosa tool¹ is adopted as the target output. A vocoder (Kong et al., 2020) is leveraged to generate the final waveform from the generated features.

Encoder-Decoder Backbone The Transformer encoder-decoder model (Vaswani et al., 2017) is used as the backbone network of SpeechT5. Please

¹<https://librosa.org/doc/latest/index.html>.

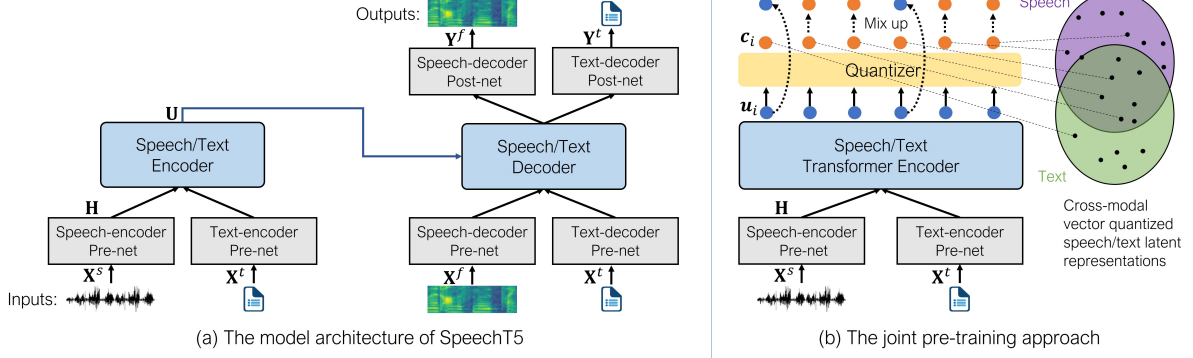


Figure 2: (a) The model architecture of SpeechT5, which contains an encoder-decoder module and six modal-specific pre/post-nets. Most spoken language processing tasks can be learned by concatenating the encoder-decoder module and the corresponding pre-net and post-net. (b) By sharing discrete tokens across modalities, the joint pre-training approach builds bridges between speech and text. Hidden states and latent units are mixed up and used as the inputs of the cross-attention module in the decoder.

refer to Vaswani et al. (2017) for more details. We employ the relative position embedding (Shaw et al., 2018) to help capture the relative position differences between elements in the input. Specifically, we only add the relative position embedding to the dot-product weights of the self-attention.

Speech Pre/Post-Net The convolutional feature extractor of `wav2vec 2.0` (Baevski et al., 2020) serves as the speech-encoder pre-net to **downsample raw waveform X^s and produce a sequence of a speech utterance $H = (h_1, \dots, h_{N_h})$** . The speech-decoder pre-net is a neural network composed of three fully connected layers with the ReLU activation, fed with the log Mel-filterbank X^f . To support multi-speaker TTS and VC, the speaker embedding extracted with the x-vector (Snyder et al., 2018) is concatenated with the output of the speech-decoder pre-net followed by a linear layer. The speech-decoder post-net consists of two modules. The first module uses a linear layer fed with the decoder output to predict the log Mel-filterbank $Y^f = (y_1^f, \dots, y_{N_f}^f)$, followed by five 1-dimensional convolutional layers to produce a residual to refine the predicted Y^f . Another linear module is added to project the decoder output to a scalar for predicting the stop token.

Text Pre/Post-Net We use shared embeddings as the text-encoder pre-net and text-decoder pre/post-nets. The pre-net transforms a token index into an embedding vector. The post-net transforms the hidden state into the probability distribution of tokens, normalized by the softmax function.

2.2 Pre-Training

The proposed SpeechT5 model can be pre-trained with large-scale collections of unlabeled speech and text corpus. The proposed joint pre-training method can align the textual and acoustic information into a unified semantic space.

Speech Pre-Training Leveraging unlabeled speech data \mathcal{D}^s to learn general speech representations for both classification and generation tasks, SpeechT5 is trained with two types of tasks: bidirectional masked prediction and sequence-to-sequence generation.

Following HuBERT (Hsu et al., 2021), the bidirectional masked prediction leverages a masked language model similar to BERT (Devlin et al., 2019) for the encoder, in which an acoustic unit discovery model provides the frame-level targets $\mathbf{Z} = (z_1, \dots, z_{N_h})^2$. Specifically, we apply span mask strategies to the output \mathbf{H} from speech-encoder pre-net, where 8% of timesteps are randomly selected as start indices, and spans of 10 steps are masked. The Transformer encoder takes masked \mathbf{H} as the input and produces hidden representations $\mathbf{U} = (u_1, \dots, u_{N_h})$. Based on these hidden representations, the cross-entropy loss is computed over masked timesteps as

$$\mathcal{L}_{mlm}^s = \sum_{n \in \mathcal{M}} \log p(z_n | \hat{\mathbf{H}}, n), \quad (1)$$

where $\hat{\mathbf{H}}$ denotes the masked version of \mathbf{H} , \mathcal{M}

²The target labels are generated by clustering outputs of the 6-th Transformer layer in the first iteration HuBERT BASE model via the k -means clustering method with 500 clusters.

denotes the set of masked timesteps, and \mathbf{z}_n denotes the frame-level target at timestep n from \mathbf{Z} .

Furthermore, we propose to **reconstruct the original speech via a sequence-to-sequence generation task**, given the randomly masked input as introduced in bidirectional masked prediction. Following seq2seq TTS models (Li et al., 2019), we enforce the corresponding predicted output \mathbf{Y}^f , which is generated through the speech-decoder pre-net, Transformer decoder, and speech-decoder post-net, to be close to the original \mathbf{X}^f by minimizing their L_1 -distance as

$$\mathcal{L}_1^s = \sum_{n=1}^{N^f} \|\mathbf{y}_n^f - \mathbf{x}_n^f\|_1, \quad (2)$$

where \mathbf{x}_n^f denotes n -th an 80-dimensional log Mel-filterbank from \mathbf{X}^f . Besides, we use the binary cross-entropy (BCE) loss \mathcal{L}_{bce}^s for the stop token.

Text Pre-Training With unlabeled text data \mathcal{D}^t , SpeechT5 is trained to reconstruct the model output $\mathbf{Y}^t = (\mathbf{y}_1^t, \dots, \mathbf{y}_{N^t}^t)$ to the original text \mathbf{X}^t , using the corrupted text $\hat{\mathbf{X}}^t = (\hat{x}_1^t, \dots, \hat{x}_M^t)$ as the input generated with a mask-based noising function. Following the text infilling approach in BART³ (Lewis et al., 2020), we randomly sample 30% of text spans to mask, where the span length of text spans draws from a Poisson distribution ($\lambda = 3.5$), and each span is replaced with a single mask token. Formally, SpeechT5, including text-encoder pre-net, encoder-decoder model, and text-decoder pre/post nets, is optimized to generate the original sequence with the maximum likelihood estimation as

$$\mathcal{L}_{mle}^t = \sum_{n=1}^{N^t} \log p(\mathbf{y}_n^t | \mathbf{y}_{<n}^t, \hat{\mathbf{X}}^t), \quad (3)$$

Joint Pre-Training The above pre-training methods can only leverage speech or text data to model acoustic or language information individually. To build a cross-modality mapping between speech and text, which is essential for tasks such as ASR and TTS, we propose a cross-modal vector quantization method to learn representations capturing the modality-invariant information.

Specifically, we utilize vector quantized embeddings as a bridge to align the speech representation and text representation through a shared codebook,

as shown in Figure 2(b). Inspired by VQ-VAE (Oord et al., 2017) and SemFace (Ren et al., 2021), we first use the quantizer to convert these continuous speech/text representations \mathbf{u}_i from the output of the encoder into discrete representations \mathbf{c}_i from a fixed-size codebook \mathbf{C}^K , which contains K learnable embeddings. Then, the nearest neighbor search is performed between the encoder output and the embedding of each latent code via the L_2 distance as

$$\mathbf{c}_i = \arg \min_{j \in [K]} \|\mathbf{u}_i - \mathbf{c}_j\|_2, \quad (4)$$

where \mathbf{c}_j is the j -th quantized vector in the codebook. Note that we do the same operation for the output of the speech and text encoders with a shared codebook.

Then, we randomly replace a proportion (10%) of the contextual representations with quantized latent representations in the corresponding time steps and calculate the cross-attention upon the mixed representations, which can explicitly guide the quantizer to utilize the cross-modal information. The diversity loss is used to encourage sharing more codes by maximizing the entropy of the averaged Softmax distribution as

$$\mathcal{L}_d = \frac{1}{K} \sum_{k=1}^K p_k \log p_k, \quad (5)$$

where p_k is the averaged probability of choosing the k -th code in the codebook.

The final pre-training loss with unlabeled speech and text data can be formulated as

$$\mathcal{L} = \mathcal{L}_{mlm}^s + \mathcal{L}_1^s + \mathcal{L}_{bce}^s + \mathcal{L}_{mle}^t + \gamma \mathcal{L}_d. \quad (6)$$

where γ is set to 0.1 during pre-training.

2.3 Fine-Tuning

After pre-training, we fine-tune the encoder-decoder backbone via the loss of the downstream task. The goal is to measure the learning abilities of SpeechT5, and we study the performance on a diverse set of downstream tasks such as ASR, TTS, ST, VC, SE, and SID. All of the spoken language processing tasks that we consider can be learned by concatenating the outputs of the encoder-decoder backbone and corresponding pre-net and post-net. Taking ASR as an example, the final model consists of the speech-encoder pre-net, encoder-decoder, text-decoder pre-net, and text-decoder post-net,

³We conducted experiments to compare the BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) mask strategies, which can be found in Appendix A.

Model	LM	dev-clean	dev-other	test-clean	test-other
wav2vec 2.0 BASE (Baevski et al., 2020)	-	6.1	13.5	6.1	13.3
HuBERT BASE (Hsu et al., 2021) †	-	5.5	13.1	5.8	13.3
Baseline (w/o CTC)	-	5.8	12.3	6.2	12.3
Baseline	-	4.9	11.7	5.0	11.9
SpeechT5 (w/o CTC)	-	5.4	10.7	5.8	10.7
SpeechT5	-	4.3	10.3	4.4	10.4
DiscreteBERT (Baevski et al., 2019)	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE (Baevski et al., 2020)	4-gram	2.7	7.9	3.4	8.0
HuBERT BASE (Hsu et al., 2021)	4-gram	2.7	7.8	3.4	8.1
wav2vec 2.0 BASE (Baevski et al., 2020)	Transf.	2.2	6.3	2.6	6.3
Baseline	Transf.	2.3	6.3	2.5	6.3
SpeechT5	Transf.	2.1	5.5	2.4	5.8

Table 1: Results of ASR (speech to text) on the LibriSpeech dev and test sets when training on the 100 hours subset of LibriSpeech. † indicates that results are not reported in the corresponding paper and evaluated by ourselves.

which are initialized by SpeechT5 and fine-tuned via the cross-entropy loss on the corresponding training data. The baseline systems have the same architecture as SpeechT5, but the weights of the baseline encoder are initialized by the HuBERT BASE model (Hsu et al., 2021) if the input data of the downstream tasks is speech. It allows raw waveform as the model input and can provide a strong baseline.

3 Experiments

3.1 Pre-Training Setup

All models are implemented in Fairseq⁴ (Ott et al., 2019). The encoder-decoder backbone contains 12 Transformer encoder blocks and 6 Transformer decoder blocks, where the model dimension is 768, the inner dimension (FFN) is 3,072, and the number of attention heads is 12. The above encoder setting is the same as that in wav2vec 2.0 BASE and HuBERT BASE. The speech-encoder pre-net contains 7 blocks of temporal convolutions, each of which is composed of 512 channels with strides (5, 2, 2, 2, 2, 2) and kernel sizes (10, 3, 3, 3, 3, 2, 2). For the speech-decoder pre-net and post-net, we use the same setting as the pre-net and post-net in Shen et al. (2018) except that the number of channels of the post-net is 256. For text-encoder/decoder pre/post-net, a shared embedding layer with dimension 768 is used. For the vector quantization, we use two codebooks with 100 entries for the shared codebook module, resulting in a theoretical maximum of $K = 10^4$ code entries.

For speech pre-training, we use the full 960 hours of LibriSpeech audio (Panayotov et al., 2015).

For text pre-training, we use the normalized language model training text of LibriSpeech as unlabeled data, which contains 400M sentences.⁵ We optimize the model with Adam (Kingma and Ba, 2014) by warming up the learning rate for the first 8% of updates to a peak of 2×10^{-4} , which is linearly decayed for the following updates. We pre-train the proposed SpeechT5 model on 32 V100 GPUs with a batch size of around 90s samples per GPU for speech and 12k tokens per GPU for text and set the update frequency to 2 for 500k steps.

3.2 Evaluation on ASR

We fine-tune the ASR model with the LibriSpeech 100/960 hours data and train the language model (LM) with the LibriSpeech LM text data, which is used for shallow fusion (Gulcehre et al., 2015) during the ASR inference. Besides the cross-entropy loss for the decoder, we add an extra linear layer to calculate the connectionist temporal classification (CTC) loss on the top of the encoder (Watanabe et al., 2017), so that we can apply the joint CTC/attention decoding (Hori et al., 2017) to boost the performance. We measure the performance of ASR by the word error rate (WER). The implementation details can be found in Appendix B.1.

The results of ASR on the 100 hours set of LibriSpeech are reported in Table 1. We compare with several state-of-the-art self-supervised approaches, including DiscreteBERT (Baevski et al., 2019), wav2vec 2.0 (Baevski et al., 2020), and HuBERT (Hsu et al., 2021). Without LM fusion, the baseline outperforms wav2vec 2.0 BASE and HuBERT BASE with the help of the joint CTC/attention decoding, which shows the importance of the decoder.

⁴<https://github.com/pytorch/fairseq>

⁵<https://www.openslr.org/11>

Model	WER		MCD	
	bdl to slt	clb to slt	bdl to slt	clb to slt
VTN w/ ASR (Huang et al., 2021)	11.1%	10.9%	6.50	6.11
VTN w/ TTS (Huang et al., 2021)	7.6%	9.1%	6.33	6.02
Many-to-many VTN (Kameoka et al., 2021)	-	-	6.13	5.97
Baseline	21.5%	10.8%	6.26	6.16
SpeechT5	7.8%	6.4%	5.93	5.87

Table 2: Results of VC (speech to speech) on the CMU Arctic. The bdl, clb, and slt denote three speakers.

The proposed SpeechT5 model achieves significant improvements on all settings compared to wav2vec 2.0 BASE, HuBERT BASE and our strong baselines, demonstrating the superiority of the proposed pre-training method. Furthermore, when decoding with LM fusion, SpeechT5 obtains the lower WERs than wav2vec 2.0 BASE on all sets and achieves the state-of-the-art performance. Due to space constraints, the results of 960h fine-tuning experiments are reported in Appendix C.

3.3 Evaluation on TTS

We fine-tune the pre-trained model on the 460-hours LibriTTS clean sets (Zen et al., 2019) with the L_1 loss, \mathcal{L}_{bce}^s loss, and attention loss (Tachibana et al., 2018). We utilize the HiFi-GAN (Kong et al., 2020) vocoder to convert the log Mel-filterbank to the raw waveform. We evaluate the Naturalness with the open-source NISQA-TTS (Mittag and Möller, 2020), the mean option score (MOS), and the comparison mean option score (CMOS) by native speakers on the randomly selected 200 sentences with various lengths (no overlapping with training data) generated by different models, in which case we keep the text content consistent. More details can be found in Appendix B.2.

Model	Naturalness	MOS	CMOS
Ground Truth	-	3.87 ± 0.04	-
Baseline	2.76	3.56 ± 0.05	0
SpeechT5	2.91	3.65 ± 0.04	+0.290

Table 3: Results of TTS (text to speech) on the LibriTTS.

Table 3 shows the experimental results of TTS. The proposed SpeechT5 trained without \mathcal{L}_{mlm}^s is considered because the bidirectional masked prediction loss is proposed to help the encoder learn to encode the speech signal, and this variant achieves superior Naturalness, as shown in Table 13 (in Appendix D). The proposed SpeechT5 model behaves

better than baseline and achieves a performance of 2.91 Naturalness and 3.65 MOS. Furthermore, our proposed SpeechT5 obtains a gain of +0.29 in CMOS with respect to the baseline model, which suggests the proposed pre-training method significantly improves the speech generation quality.

3.4 Evaluation on ST

We evaluate the ST task on the MUST-C dataset (Di Gangi et al., 2019), including English-German (EN-DE) and English-French (EN-FR) translation tasks. We use the default training setting of speech translation in Fairseq ST (Wang et al., 2020), and we also average the last 10 checkpoints and use a beam size of 5 for decoding. Translation results are evaluated with case-sensitive BLEU (Papineni et al., 2002). Details about the dataset and fine-tune setting are introduced in Appendix B.3.

Model	EN-DE	EN-FR
Fairseq ST (Wang et al., 2020)	22.70	32.90
ESPnet ST (Inaguma et al., 2020)	22.91	32.69
Adapter Tuning (Le et al., 2021)	24.63	34.98
Baseline	23.43	33.76
SpeechT5 (w/o initializing decoder)	24.44	34.53
SpeechT5	25.18	35.30

Table 4: Results of ST (speech to text) on the MUST-C EN-DE and EN-FR.

We list the BLEU scores of ST in Table 4. The result of SpeechT5 without initializing the decoder is also reported since we do not pre-train the decoder with German or French data, and it outperforms the strong baseline whose encoder is initialized by HuBERT encoder. The proposed SpeechT5 further beats the SpeechT5 without initializing the decoder, and achieves a significant improvement of 1.75 and 1.54 BLEU scores than baseline in EN-DE and EN-FR tasks, respectively, which demonstrates the effectiveness and superiority of our method. Besides, our SpeechT5 model outperforms existing models such as Fairseq ST (Wang et al., 2020),

ESPnet ST (Inaguma et al., 2020), and Adapter Tuning (Le et al., 2021) that employs adapter modules to be further specialized in each language pair from different pre-trained models.

3.5 Evaluation on VC

VC aims to convert a speaker-dependent source speech waveform into a different one while preserving linguistic information of the source speech waveform. We follow the many-to-many setting and utilize speech recordings of four speakers in the CMU Arctic (Kominek and Black, 2004), including clb, bdl, slt, and rms. For the waveform synthesis, we use the Parallel WaveGAN (Yamamoto et al., 2020), a non-autoregressive variant of the WaveNet vocoder. We employ the average of MCD (Mel-Cepstral Distortion) and WER as the metrics for the VC task. More details about the dataset and fine-tune setting are given in Appendix B.4.

We show the results of VC in Table 2, where we list the conversion from speaker bdl to slt and clb to slt as used in the voice Transformer network (VTN) (Huang et al., 2021). The experimental results demonstrate that the proposed SpeechT5 model achieves a significant gain than the strong baseline model. The proposed SpeechT5 model also outperforms the state-of-the-art VTN variants in terms of MCD, including VTN fine-tuned from ASR or TTS (Huang et al., 2021) and many-to-many VTN (Kameoka et al., 2021).

3.6 Evaluation on SE

SE is the task of removing background noise from a degraded speech signal and improving the intelligibility and the perceived quality of the signal. We use the WSJ0 Hipster Ambient Mixtures (WHAM!) dataset (Wichern et al., 2019) and conduct the 16 kHz max enhance-single task that recovers the signal from a mixture of only the first WSJ0 speaker and noise. We utilize HiFi-GAN to transform the log Mel-filterbank to the raw waveform. Since the input and output lengths are probably different in the encoder-decoder model, we can not evaluate it by PESQ (Rix et al., 2001) and ESTOI (Jensen and Taal, 2016), so we evaluate the negative impact on the ASR performance by WER. The implementation details of SE are in Appendix B.5.

As shown in Table 5, our strong baseline model recovers contents from the noisy speech, achieving 10.9% WER from 76.1% WER. Moreover, the proposed SpeechT5 model gets a relative 9% WER reduction compared to the strong baseline model.

Model	WER
Ground Truth Speech	3.2%
Noisy Speech (Wichern et al., 2019)	76.1%
Baseline	10.9%
SpeechT5	8.9%

Table 5: Results of SE (speech to speech) on the WHAM!.

The results suggest that although the noisy speech with WHAM! is challenging as summarized in Table 12 (in Appendix B.5), the proposed encoder-decoder framework can effectively suppress the noise and recover the content.

3.7 Evaluation on SID

We convert SID, a multi-class classification task of classifying each utterance for its speaker identity, to a speech to text task by sequence to sequence model. Compared to the ASR task, the text embedding table is replaced by a speaker embedding table, and the decoder predicts speaker identifies at the first step. We adopt the VoxCeleb1 dataset (Nagrani et al., 2017), which contains over 100,000 speech records uttered by 1,251 celebrities extracted from videos uploaded to YouTube. The top-1 speaker classification accuracy (ACC) is used as the evaluation metric of SID. Refer to Appendix B.6 for more details about the dataset and fine-tuning.

Model	ACC
SUPERB (Yang et al., 2021)	
wav2vec 2.0 BASE (Baevski et al., 2020)	75.18%
HuBERT BASE (Hsu et al., 2021)	81.42%
HuBERT LARGE (Hsu et al., 2021)	90.33%
SpeechNet (Chen et al., 2021b)	
Single Task	86.00%
Multi-Task with TTS	87.90%
Thin ResNet-34 (Chung et al., 2020)	89.00%
Ours	
Baseline	91.92%
SpeechT5	96.49%

Table 6: Results of SID (speech to text) on the VoxCeleb1. The SUPERB fine-tuning freezes the encoder.

As shown in Table 6, our baseline is superior to existing Transformer-based methods such as SpeechNet (Chen et al., 2021b) and pre-trained models from SUPERB (Yang et al., 2021). Moreover, it outperforms ResNet-based architectures such as Thin ResNet-34 (Chung et al., 2020), indicating the superiority of the encoder-decoder ar-

chitecture for the SID task. The SpeechT5 further improves the performance compared to baseline and achieves the state-of-the-art performance (i.e., 96.49% accuracy), which demonstrates the effectiveness of the proposed pre-training technique.

3.8 Ablation Study

To better understand why the proposed SpeechT5 model is effective, we **investigate the influence of the pre-training methods** by removing each of them independently.

Model	ASR		VC	SID
	clean	other		
SpeechT5	4.4	10.7	5.93	96.49%
w/o Speech PT	-	-	6.49	38.61%
w/o Text PT	5.4	12.8	6.03	95.60%
w/o Joint PT	4.6	11.3	6.18	95.54%
w/o \mathcal{L}_{mlm}^s	7.6	22.4	6.29	90.91%

Table 7: Ablation study for the SpeechT5 model. Different variants of the SpeechT5 model, including the SpeechT5 model without speech pre-training (PT), text pre-training, joint pre-training method, or the bidirectional masked prediction loss, are evaluated on the ASR (test subsets with WER), VC (bdl to slt with MCD), and SID (test set with ACC) tasks.

As shown in Table 7, we can draw the following conclusions: (1) The pre-training methods, including speech pre-training, text pre-training, and joint pre-training method, are important to SpeechT5 since without each of them, the performance of all tasks will degrade significantly; (2) Speech pre-training is more critical than text pre-training on these tasks that need to encode speech, and the ASR model fine-tuned from SpeechT5 without speech pre-training even can not converge; (3) Without the joint pre-training method, the performance of the ASR model decreases, which demonstrates that the learned alignment from joint pre-training brings benefits for cross-modality tasks; (4) The masked language model learning \mathcal{L}_{mlm}^s of speech data is mainly responsible for extracting acoustic features and learning better speech representation, which is beneficial to ASR and SID tasks.

4 Related Work

Large-scale pre-training models such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), wav2vec 2.0 (Baevski et al., 2020), and HuBERT (Hsu et al., 2021) have drawn much attention in the NLP and speech communities, due to its strong capabil-

ity of generalization and efficient usage of large-scale data (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019; Lewis et al., 2020; Chen et al., 2021c; Baevski et al., 2020; Lakhota et al., 2021; Kharitonov et al., 2021; Chen et al., 2021a). However, the research mentioned above effects gear towards single-modal learning, hence they can only be used in either text or speech modeling. Although some speech-language pre-training work (Chung et al., 2021b; Kim et al., 2021; Qian et al., 2021) attempts to improve spoken language understanding tasks, these methods only focus on an encoder with task-specific layers for different tasks and do not pre-train a decoder for generation tasks such as speech synthesis or text generation. Besides, a series of research work begins to investigate joint text and speech training (Han et al., 2021; Ye et al., 2021; Tang et al., 2021a; Zheng et al., 2021; Tang et al., 2021b), but they are mainly designed for speech to text tasks.

The proposed SpeechT5 method is most related to T5 (Raffel et al., 2020). The core idea of the T5 model, a unified framework for a variety of text-based language problems, is to treat every text processing problem as a “text-to-text” problem. SpeechT5 is also related to Speech Chain (Tjandra et al., 2020), which leverages the ASR model and TTS model to build a closed-loop machine speech chain to train models on the concatenation of both labeled and unlabeled data, and SpeechNet (Chen et al., 2021b), which designs a universal modularized model to perform multiple speech processing tasks with multi-task learning. The differences from the above models are that (1) SpeechT5 is a shared cross-modal encoder-decoder framework, whose input and output are speech or text through multiple pre/post-nets; (2) SpeechT5 attempts to pre-train and improve the universal model with large-scale unlabeled text and speech data.

Another related work is SUPERB (Yang et al., 2021), a benchmark to examine the capability of pre-trained models such as HuBERT (Hsu et al., 2021). SUPERB focuses on investigating a simple framework to learn SUPERB tasks with a frozen and shared pre-trained encoder and lightweight prediction modules fine-tuned for each task. In contrast, the goal of SpeechT5 is to learn all spoken language processing tasks by fine-tuning a unified-modal encoder-decoder model, which is pre-trained on unlabeled speech and text corpus.

5 Conclusion

In this paper, we have proposed SpeechT5 as a pre-trained encoder-decoder model for various spoken language processing tasks. We convert all spoken language processing tasks into a speech/text to speech/text format and propose a novel joint pre-training method to utilize cross-modal information by leveraging the unlabeled speech and text data. The proposed unified encoder-decoder model can support generation tasks such as speech translation and voice conversion. Massive experiments show that SpeechT5 significantly outperforms all baselines in several spoken language processing tasks. In the future, we are going to pre-train the SpeechT5 with a larger model and more unlabeled data. We are also interested in extending the proposed SpeechT5 framework to address multilingual spoken language processing tasks for future work.

Acknowledgments

We thank Yanqing Liu and Sheng Zhao for their help in TTS human evaluation. We also want to thank the anonymous reviewers for insightful comments and suggestions.

References

- Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. [Effectiveness of self-supervised pre-training for speech recognition](#). *arXiv preprint arXiv:1911.03912*.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). In *Proceedings of the 34th Conference on Neural Information Processing Systems*, volume 33, pages 12449–12460.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021a. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *arXiv preprint arXiv:2110.13900*.
- Yi-Chen Chen, Po-Han Chi, Shu-wen Yang, Kai-Wei Chang, Jheng-hao Lin, Sung-Feng Huang, Da-Rong Liu, Chi-Liang Liu, Cheng-Kuang Lee, and Hung-yi Lee. 2021b. [Speechnet: A universal modularized model for speech processing tasks](#). *arXiv preprint arXiv:2105.03070*.
- Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Gary Wang, and Pedro Moreno. 2021c. [Injecting text in self-supervised speech pre-training](#). *arXiv preprint arXiv:2108.12226*.
- Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin shan Lee. 2020. [SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering](#). In *Proceedings of Interspeech 2020*, pages 4168–4172.
- Joon Son Chung, Jaesung Huh, and Seongkyu Mun. 2020. [Delving into VoxCeleb: Environment invariant speaker recognition](#). In *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 349–356.
- Yu-An Chung and James Glass. 2018. [Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech](#). In *Proceedings of Interspeech 2018*, pages 811–815.
- Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021a. [w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training](#). In *Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.
- Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2021b. [SPLAT: Speech-language joint pre-training for spoken language understanding](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1897–1907.
- Alexis CONNEAU and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, volume 32.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. [MuST-C: a Multilingual Speech Translation Corpus](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017.
- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. [Unified language model pre-training for natural language understanding and generation](#). In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, volume 32, pages 13063–13075.

- Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Hui-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. [On using monolingual corpora in neural machine translation](#). *arXiv preprint arXiv:1503.03535*.
- Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Proceedings of the 2021 Findings of the Association for Computational Linguistics*, pages 2214–2225.
- Takaaki Hori, Shinji Watanabe, and John Hershey. 2017. [Joint CTC/attention decoding for end-to-end speech recognition](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [Hubert: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, and Tomoki Toda. 2021. [Pretraining techniques for sequence-to-sequence voice conversion](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:745–755.
- Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. [Espnet-st: All-in-one speech translation toolkit](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311.
- Jesper Jensen and Cees H. Taal. 2016. [An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers](#). *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(11):2009–2022.
- Hirokazu Kameoka, Wen-Chin Huang, Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Tomoki Toda. 2021. [Many-to-many voice transformer network](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:656–670.
- Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.
- Minjeong Kim, Gyuwan Kim, Sang-Woo Lee, and Jung-Woo Ha. 2021. [St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding](#). In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7478–7482.
- Diederik P Kingma and Jimmy Ba. 2014. [Adam: A method for stochastic optimization](#). *arXiv preprint arXiv:1412.6980*.
- John Kominek and Alan W Black. 2004. [The cmu arctic speech databases](#). In *Proceedings of the Fifth ISCA workshop on speech synthesis*.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. [Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis](#). In *Proceedings of the 34th Conference on Neural Information Processing Systems*, volume 33, pages 17022–17033.
- Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.
- Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. [Lightweight adapter tuning for multilingual speech translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. [Neural speech synthesis with transformer network](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6706–6713.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Gabriel Mittag and Sebastian Möller. 2020. [Deep learning based assessment of synthetic speech naturalness](#). In *Proceedings of Interspeech 2020*, pages 1748–1752.
- Arsha Nagrani, Joon Son Chung, and Andrew Senior. 2017. Voxceleb: A large-scale speaker identification dataset. In *Proceedings of the Interspeech 2017*, pages 2616–2620.
- Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30.

- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.
- Yao Qian, Ximo Bian, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. 2021. Speech-language pre-training for end-to-end spoken language understanding. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7458–7462.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Chandan K. A. Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. 2021. ICASSP 2021 deep noise suppression challenge. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2021-June, pages 6623–6627.
- Shuo Ren, Long Zhou, Shujie Liu, Furu Wei, Ming Zhou, and Shuai Ma. 2021. Semface: Pre-training encoder and decoder with a semantic interface for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4518–4527.
- A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. 2001. Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752.
- Braun Sebastian and Tashev Ivan. 2020. Data augmentation and loss normalization for deep noise suppression. In *Proceedings of Speech and Computer*, pages 79–86.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.
- Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4779–4783.
- David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5329–5333.
- Xingchen Song, Guangsen Wang, Zhiyong Wu, Yiheng Huang, Dan Su, Dong Yu, and Helen Meng. 2019. Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks. In *Proceedings of the Interspeech 2020*, pages 3765–3769.
- Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2020. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.
- Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4784–4788.
- Yun Tang, Juan Pino, Xian Li, Changan Wang, and Dmitriy Genzel. 2021a. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4252–4261.

- Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitry Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6209–6213.
- Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. [Machine speech chain](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30, pages 6000–6010.
- Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39.
- Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumataani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *Proceedings of the 2021 International Conference on Machine Learning*, pages 10937–10947.
- Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Proceedings of the Interspeech 2018*, pages 2207–2211.
- Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. [Hybrid ctc/attention architecture for end-to-end speech recognition](#). *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.
- Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. WHAM!: Extending speech separation to noisy environments. In *Proceedings of Interspeech 2019*, pages 1368–1372.
- Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. [Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram](#). In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6199–6203.
- Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. [Superb: Speech processing universal performance benchmark](#). *arXiv preprint arXiv:2105.01051*.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 32.
- Rong Ye, Mingxuan Wang, and Lei Li. 2021. [End-to-End Speech Translation via Cross-Modal Progressive Training](#). In *Proceedings of the Interspeech 2021*, pages 2267–2271.
- Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proceedings of the Interspeech 2019*, pages 1526–1530.
- Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proceedings of the 2021 International Conference on Machine Learning*, pages 12736–12746.

A Comparisons of Text Mask Strategies

We compare the performance when using the BART (Lewis et al., 2020) or T5 (Raffel et al., 2020) strategies for text masking on the ASR task, as reported in Table 10. The BART strategy achieves comparable or better performance than the T5 strategy under different inference settings.

B Implementation Details

B.1 ASR

Dataset We use the LibriSpeech corpus and fine-tune on two labeled data settings: 960 hours of transcribed Librispeech and the train-clean-100 subset comprising 100 hours (100 hours labeled). We train the language model by the LibriSpeech language model (LM) text data, which is used for shallow fusion (Gulcehre et al., 2015) during the ASR inference.

Fine-Tuning Details We fine-tune the model with the CTC loss and the cross-entropy loss, where the loss weights are 0.5 for both of them. We train on 8 V100 GPUs with a batch size of up to 256k audio samples per GPU. The learning rate is warmed up for the first 10% steps, hold as a constant for the following 40% steps, and is decayed linearly for the rest steps. Table 8 summarizes the hyperparameters for ASR experiments of 100 hours and 960 hours sets.

Hyperparameter	100 hours	960 hours
updates	80k	320k
learning rate	6e-5	1.3e-4
time-step mask prob.	0.075	0.05
channel mask prob	0.008	0.0016

Table 8: The setting of hyperparameters for ASR fine-tuning.

Language Model and Decoding We train a character-level LM for the ASR inference. The model has the same architecture as the Transformer LM in Synnaeve et al. (2020), which is used for decoding of wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). The LM contains 20 blocks of Transformer decoder with the model dimension of 1280, inner dimension of 6144, and 16 attention heads. To investigate the difference of the performance between our LM and the LM in Synnaeve et al. (2020), we evaluate the word-level

Language Model	dev	
	clean	other
Word 4-gram (Synnaeve et al., 2020)	148.0	136.6
Word Transf. (Synnaeve et al., 2020)	48.2	50.2
Character Transf.	56.5	59.3

Table 9: Word-level perplexities of language models on dev-clean/other sets of LibriSpeech.

perplexities of these LMs on the LibriSpeech dev-clean/other sets as shown in Table 9. The Transformer LM used for SpeechT5 gets 56.5 perplexity on the dev-clean set and 59.3 perplexity on the dev-other set, which are higher than the perplexities of word Transformer LM in Synnaeve et al. (2020). It suggests that we may achieve better performance on the ASR task if the perplexities of our LM are similar to the LM in Synnaeve et al. (2020).

During decoding, the beam size is set to 30 for all experiments. We select the model with the highest accuracy on dev-other set for inference and apply the joint CTC/attention decoding (Hori et al., 2017) to further improve the performance. The model generates the output transcription by the beam search algorithm, which aims to maximize

$$\alpha \log P_{Dec} + (1 - \alpha) \log P_{CTC} + \beta \log P_{LM} \quad (7)$$

where α and β are weights for the log probabilities, P_{Dec} , P_{CTC} , and P_{LM} are the probabilities of the decoder, CTC, and LM, respectively. We set α to 0.5 and β to 1.0 for fine-tuning experiments of 100 hours set, and set α to 0.9 and β to 0.7 for fine-tuning experiments of 960 hours set.

B.2 TTS

Dataset and Evaluation Metrics We use the 460-hours LibriTTS clean sets (Zen et al., 2019), a multispeaker corpus of read English speech from the audiobooks of the LibriVox project, as TTS training dataset. We trim the waveform as ESPnet recipe (Watanabe et al., 2018). The WER is evaluated by using the open-source ASR model wav2vec 2.0 CTC⁶. The naturalness of synthetic speech is estimated by using the open-source TTS naturalness prediction model NISQA-TTS⁷ (Mittag and Möller, 2020).

Fine-Tuning Details Besides the L_1 loss and BCE loss, we add an additional attention loss

⁶<https://huggingface.co/facebook/wav2vec2-base-960h>

⁷<https://github.com/gabrielmittag/NISQA>

Mask Strategies	CTC	LM	dev		test	
			clean	other	clean	other
BART (Lewis et al., 2020)	-	-	5.4	10.7	5.8	10.7
	✓	-	4.3	10.3	4.4	10.4
	✓	✓	2.1	5.5	2.4	5.8
T5 (Raffel et al., 2020)	-	-	5.4	11.3	5.7	11.3
	✓	-	4.3	10.7	4.4	10.7
	✓	✓	2.3	5.8	2.3	5.8

Table 10: Comparisons of mask strategies for the text pre-training under different inference settings. Models are pre-trained on the 960 hours speech data of LibriSpeech and 400M text sentences of LibriSpeech-LM corpus, and fine-tuned on the 100 hours labeled data of LibriSpeech. CTC and LM mean the Joint CTC/attention decoding (Hori et al., 2017), and language model fusion, respectively.

(Tachibana et al., 2018) to speed up model convergence. We train on 8 V100 GPUs in a speaker-independent manner by using the training data of the LibriTTS. The model is updated for 120k steps with a learning rate of 0.0004, while each GPU processes up to 45,000 tokens for a batch. The learning rate is warmed up for the first 10k steps and decayed in an inverse square root manner for the rest steps.

B.3 ST

Dataset and Evaluation Metrics We evaluate the ST task on the MUST-C dataset (Di Gangi et al., 2019), including English-German (EN-DE) and English-French (EN-FR) translation tasks. The EN-DE/EN-FR language pair consists of 408/492 hours of speech data aligned with 234K/280K translated sentences. We report the results on EN-DE and EN-FR tst-COMMON set (2641 and 2632 utterances). Translation results are computed with case-sensitive BLEU (Papineni et al., 2002).

Fine-Tuning Details ST translates speech signals in a language to text in another language. We use raw audio as speech inputs in our experiments. The training setting is the same as that in S2T model in Fairseq. We set training steps to 80K and warm-up steps to 10K. Baseline and SpeechT5 models are trained with 8 GPUs via Adam optimizer. We use 8K unigram vocabulary for both EN-DE and EN-FR. Following Fairseq ST (Wang et al., 2020), we average the last 10 checkpoints and use a beam size of 5 for decoding.

B.4 VC

Dataset and Evaluation Metrics We consider the many-to-many setting for the CMU Arctic (Kominek and Black, 2004), which contains speech recordings of four speakers, such as clb (female),

bdl (male), slt (female), and rms (male), who read the same 1,132 phonetically balanced English utterances. Thus, there are twelve different combinations of source and target speakers. For each speaker, the first 932, the last 100, and the rest 100 sentences of the 1,132 sentences are used for training, test, and validation as (Huang et al., 2021), respectively. The average of MCD is estimated by using the DTW (dynamic time warping) path between the output and ground-truth Mel-cepstra. A smaller MCD indicates better performance. The WER is evaluated by using the public ASR model HuBERT LARGE⁸, where the WER of the test set with this ASR model is comparable to that of VTN (Huang et al., 2021).

Fine-Tuning Details Besides the L_1 loss and BCE loss, we add an additional attention loss (Tachibana et al., 2018) to speed up the model convergence. The model is trained on 8 V100 GPUs by the Adam optimizer with a batch size of 20000 tokens per GPU. We assign the learning rate based on the inverse square root with the maximum learning rate of 10^{-4} within 60k steps and apply 6k warm-up steps.

B.5 SE

Dataset and Evaluation Metrics We aim to recover the content of signals contaminated by various noises and reduce the negative impact on the performance of ASR systems. The 16 kHz enhance-single task of the WHAM! dataset (Wichern et al., 2019) is used as the SE dataset. It contains 20,000 training utterances, 5,000 validation utterances, and 3,000 test utterances, where the input waveform is a mixture of only the first WSJ0⁹ speaker and noise.

⁸<https://huggingface.co/facebook/hubert-xlarge-ls960-ft>

⁹<https://catalog.ldc.upenn.edu/LDC93S6A>

Model	LM	dev-clean	dev-other	test-clean	test-other
wav2vec 2.0 BASE (Baevski et al., 2020)	-	3.2	8.9	3.4	8.5
Baseline (w/o CTC)	-	3.1	7.8	3.1	7.6
Baseline	-	2.8	7.6	2.8	7.4
SpeechT5 (w/o CTC)	-	2.8	7.6	3.1	7.3
SpeechT5	-	2.5	7.4	2.7	7.1
wav2vec 2.0 BASE (Baevski et al., 2020)	4-gram	2.0	5.9	2.6	6.1
wav2vec 2.0 BASE (Baevski et al., 2020)	Transf.	1.8	4.7	2.1	4.8
Baseline	Transf.	2.0	4.5	1.9	4.5
SpeechT5	Transf.	1.8	4.3	1.9	4.4

Table 11: WER of ASR when training on the 960 hours labeled data of LibriSpeech.

Metric	WHAM!
PESQ	1.12
ESTOI	0.48
WER (NSNet2 (Sebastian and Ivan, 2020))	45.8%

Table 12: Results of noisy speech utterances on the test set in terms of PESQ, ESTOI, and WER.

We trim the noisy segment without contents. The WER is evaluated by using the open-source ASR model¹⁰ because lengths of inputs and outputs are probably different in the encoder-decoder model. Since lengths of noisy speech utterances are the same as lengths of clean utterances, we measure the test set via speech quality (PESQ) (Rix et al., 2001), extended short-time objective intelligibility (ESTOI) (Jensen and Taal, 2016), and WER to quantify the difficulty of noisy speech, as shown in Table 12. NSNet2 is the baseline model on the 2020 Deep Noise Suppression (DNS) challenge (Reddy et al., 2021) and obtains WER of 45.8%, probably due to the mismatch between the noise intensity of the WHAM! and DNS corpus.

Fine-Tuning Details We employ the loss function as used in the fine-tuning of the VC task. The model is trained on 8 V100 GPUs by the Adam optimizer with a batch size of 16000 tokens per GPU. We assign the learning rate based on the inverse square root with the maximum learning rate 10^{-4} within 100k steps and apply 10k warm-up steps.

B.6 SID

Dataset and Evaluation Metrics We use the official split of the VoxCeleb1 dataset (Nagrani et al., 2017) for the SID task, where the test set contains 8,251 utterances from these 1,251 celebrities. The capability of identifying speakers is assessed by

classifying an utterance into the ground-truth category. Specifically, the whole utterance is taken as an input to the model to determine the speaker identity.

Fine-Tuning Details We use the cross-entropy loss and fine-tune all models on 8 V100 GPUs by the Adam optimizer with a batch size of 64 segments per GPU and the inputs of 3 seconds. The learning rate is set based on one cycle of a triangular cyclical schedule between 10^{-8} and 5×10^{-4} in 60k steps. We initialize the weights of the text embeddings layer because there are no overlapping text tokens between the vocabularies during the pre-training and the SID fine-tuning.

C Results for 960 Hours Set of LibriSpeech

We also fine-tune the model on the 960 hours set of LibriSpeech, as reported in Table 11. Experiments show that the proposed SpeechT5 model achieves significant improvement even without LM fusion, and it performs comparable or even better than wav2vec 2.0 with LM fusion.

D Results of the SpeechT5 without \mathcal{L}_{mlm}^s on the TTS task

Model	Naturalness
SpeechT5	2.79
w/o \mathcal{L}_{mlm}^s	2.91

Table 13: Comparisons between SpeechT5 and its variant without using \mathcal{L}_{mlm}^s .

We use the automatic evaluation tool NISQA-TTS to verify the performance of TTS results here, because it is convenient and cheap compared with MOS and CMOS, which need to be evaluated by humans. As shown in Table 13, the variant of

¹⁰<https://doi.org/10.5281/zenodo.4243201>

SpeechT5 trained without the loss \mathcal{L}_{mlm}^s achieves an improvement in terms of naturalness when compared with the SpeechT5. It suggests that the pre-training without the speech-specific loss brings a significant gain. Thus, we select the SpeechT5 without the loss \mathcal{L}_{mlm}^s for MOS and CMOS evaluations.